Journal of Computer-Aided Molecular Design (2018) 32:401–414
https://doi.org/10.1007/s10822-018-0101-6

Assessment of tautomer distribution using the condensed reaction graph approach

T. R. Gimadiev[1,2] · T. I. Madzhidov[2] · R. I. Nugmanov[2] · I. I. Baskin[2,3] · I. S. Antipin[2] · A. Varnek[1]

Received: 20 September 2017 / Accepted: 18 January 2018 / Published online: 29 January 2018

Abstract
We report the first direct QSPR modeling of equilibrium constants of tautomeric transformations ($\log K_T$) in different solvents and at different temperatures, which do not require intermediate assessment of acidity (basicity) constants for all tautomeric forms. The key step of the modeling consisted in the merging of two tautomers in one sole molecular graph ("condensed reaction graph") which enables to compute molecular descriptors characterizing entire equilibrium. The support vector regression method was used to build the models. The training set consisted of 785 transformations belonging to 11 types of tautomeric reactions with equilibrium constants measured in different solvents and at different temperatures. The models obtained perform well both in cross-validation ($Q^2 = 0.81$ RMSE = 0.7 $\log K_T$ units) and on two external test sets. Benchmarking studies demonstrate that our models outperform results obtained with DFT B3LYP/6-311++G(d,p) and ChemAxon Tautomerizer applicable only in water at room temperature.

Keywords QSPR · Support vector regression · Condensed graphs of reaction · Tautomerism

Introduction

Handling tautomers is a real challenge in cheminformatics [1–7]. The main problem is related to the fact that different tautomers of the same chemical compound are described by different descriptor vectors, which may affect their registration in chemical databases, the results of similarity searching, building and application of quantitative structure–activity/property relationships (QSAR/QSPR) models for any physico-chemical or biological property. According to the estimations by Sitzmann et al. [8], tautomerism is possible for more than 2/3 of unique structures in the Chemical Structure DataBase (CSDB) of the National Cancer Institute containing some 103.5 million structure records. In order to handle tautomers in chemical databases, some enumeration rules were suggested [8–10]. However, duplicates corresponding to different tautomers are still present in public and proprietary databases [11, 12]. The canonicalization rules are implemented in free or commercial tools [13–21] capable of enumerating all possible tautomers corresponding to a given chemical structure.

Different scenarios of accounting for tautomerism in the modeling studies have been reported in the literature. In the case of structure-based drug design, several tautomers with energy values within a given energetic window are considered. However the relative stability of tautomers, which is usually estimated using quantum chemical [22, 23] or force field methods [24, 25], is rarely taken into account in scoring functions. Typically, descriptors used in QSAR/QSPR studies or in similarity-based virtual screening are calculated for the canonical tautomeric form, although some efforts to account for tautomeric equilibrium have been recently reported. Thus, fuzzy topological pharmacophoric triplets [26, 27] accounting for tautomers' populations are a part of the ISIDA descriptors [28], which were largely used in structure–activity modeling and virtual screening [29–31].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10822-018-0101-6) contains supplementary material, which is available to authorized users.

✉ A. Varnek
varnek@unistra.fr

[1] University of Strasbourg, 1 rue Blaise Pascal, 67000 Strasbourg, France

[2] Kazan Federal University, 18 Kremlyovskaya Str., Kazan, Russian Federation 420008

[3] Department of Physics, Moscow State University, Moscow, Russian Federation 119991